



A Tutorial on Domain Generalization

On the Robustness of ChatGPT (and NLP Foundation Models)

An Adversarial and Out-of-distribution Perspective

Jindong Wang
Microsoft Research Asia
<https://jd92.wang/>

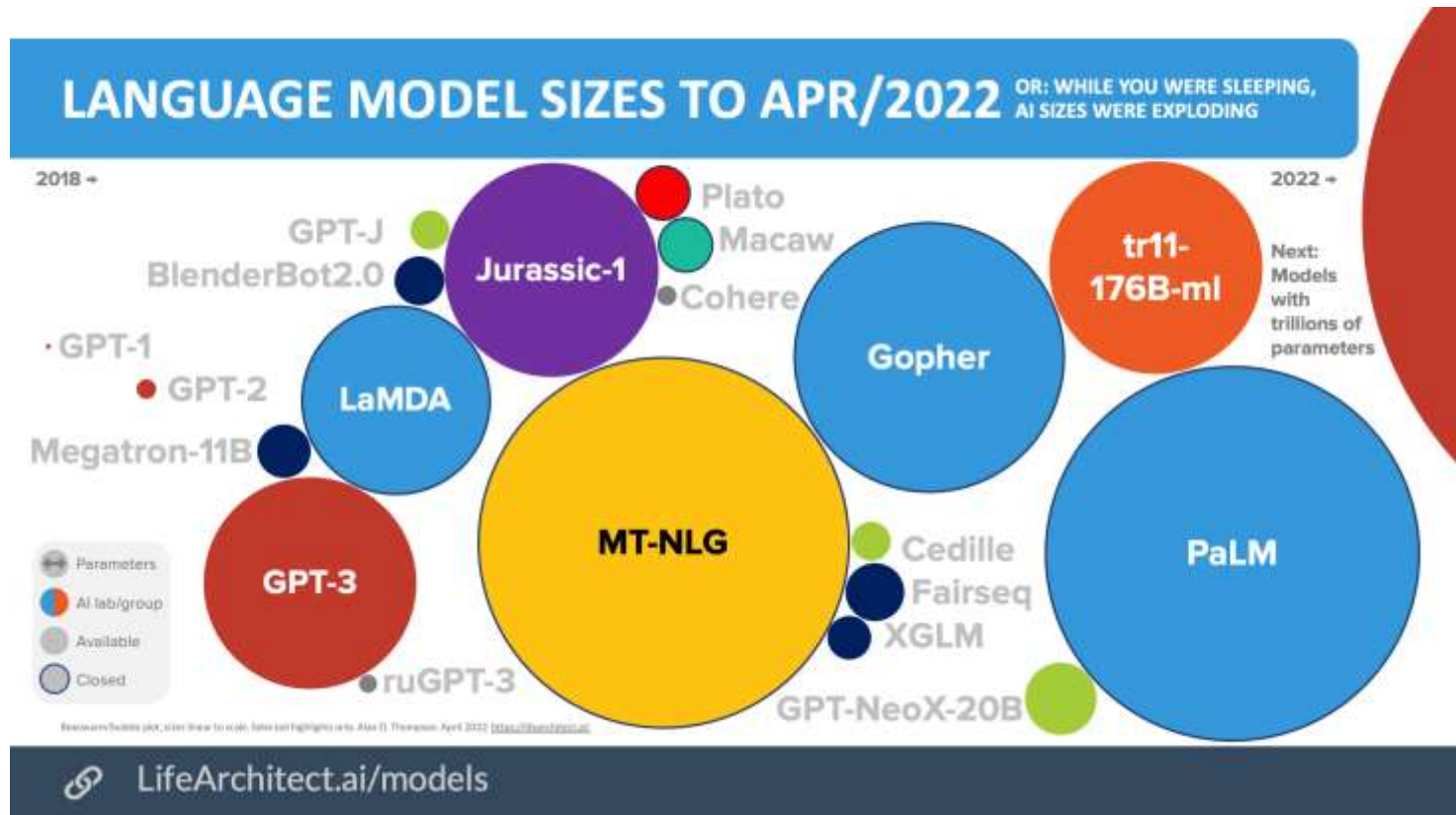
Background

- The disruption of AIGC (AI-generated content)



Behind all these technologies...

- There are large language models (LLMs), or foundation models



Year	Model	# of Parameters	Dataset Size	
Google	2019	BERT [39]	3.4E+08	16GB
HuggingFace	2019	DistilBERT [113]	6.60E+07	16GB
Google	2019	ALBERT [70]	2.23E+08	16GB
CMU & Google	2019	XLNet (Large) [150]	3.40E+08	126GB
Baidu	2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
Facebook	2019	RoBERTa (Large) [74]	3.55E+08	161GB
NVIDIA	2019	MegatronLM [122]	8.30E+09	174GB
Google	2020	T5-11B [107]	1.10E+10	745GB
Microsoft	2020	T-NLG [112]	1.70E+10	174GB
OpenAI	2020	GPT-3 [25]	1.75E+11	570GB
Google	2020	GShard [73]	6.00E+11	-
Google	2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

Superior performance

- Evaluation on ChatGPT
 - Logical reasoning (<https://arxiv.org/abs/2302.04023v1>)
 - Question answering (<https://arxiv.org/abs/2302.06476>)
 - Sequence tagging (<https://arxiv.org/abs/2302.06476>)
 - Other NLP tasks (<https://arxiv.org/abs/2302.04023v1>)
- Other concerns
 - Education (<https://arxiv.org/abs/2212.09292>)
 - Ethics (<https://arxiv.org/abs/2301.12867>)
 - Medical text (<https://arxiv.org/abs/2212.14882>)
 - Law (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335905)

Now, let's go back to domain/OOD
generalization

Is OOD/domain generalization solved by
large language models?

DG/OOD theorem

$$\epsilon^t(h) \leq \epsilon^s(h) + d_{\mathcal{H}\Delta\mathcal{H}}(P_X^s, P_X^t) + \lambda_{\mathcal{H}}$$

Target risk

Source risk

Complexity of \mathcal{H}

Source-target distribution divergence

A preliminary study

On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective

Jindong Wang^{1*}, Xixu Hu^{1,2‡†}, Wenxin Hou^{3†}, Hao Chen⁴, Runkai Zheng^{1,5‡}, Yidong Wang⁶, Linyi Yang⁷, Wei Ye⁶, Haojun Huang³, Xiubo Geng³, Binxing Jiao³, Yue Zhang⁷, Xing Xie¹

¹Microsoft Research, ²City University of Hong Kong, ³Microsoft STCA, ⁴Carnegie Mellon University, ⁵Chinese University of Hong Kong (Shenzhen), ⁶Peking University, ⁷Westlake University

<https://github.com/microsoft/robustlearn>

<https://arxiv.org/abs/2302.12095>

What do we do?

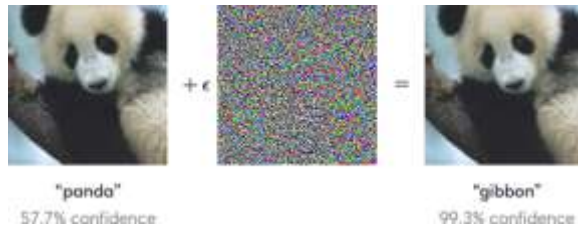
- Task: evaluating the OOD and adv. Robustness of ChatGPT
- Data:
 - OOD: use the Flipkart and DDX dataset
 - Adv: use the AdvGLUE dataset
- How:
 - Query the web from ChatGPT and gather the answer
 - Considering comparisons with other foundation models
- Prompt:
 - Design prompts for ChatGPT for our task

<https://chat.openai.com/chat>

A recap on adversarial robustness

- Adv. Examples:

- Resilience to imperceptible perturbations



$$\min_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \in \mathcal{D}} \max_{|\delta| \leq \epsilon} \ell[f(\mathbf{x} + \delta), y]$$

- OOD generalization:

- Resilience to distribution shift



Adv. and OOD robustness are closely related: both are input perturbations

Adv. dataset

- AdvGLUE

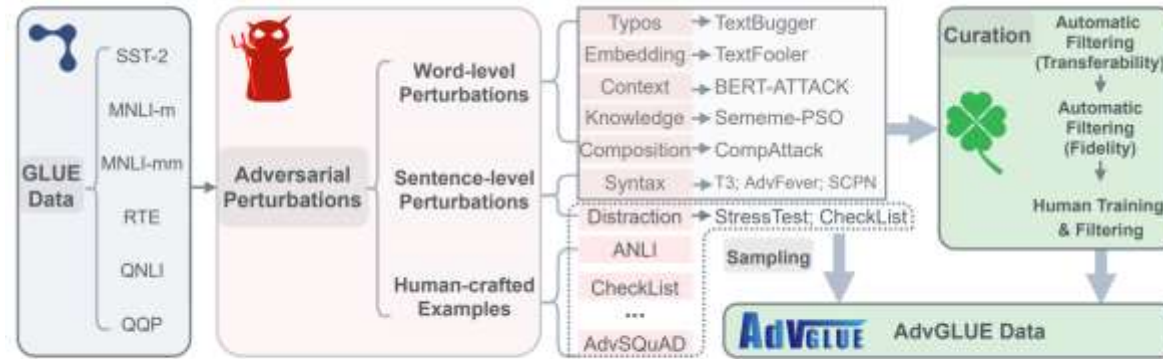


Figure 1: Overview of the AdvGLUE dataset construction pipeline.

Table 1: **Statistics of AdvGLUE benchmark.** We apply *all* word-level perturbations (C1=*Embedding-similarity*, C2=*Typos*, C3=*Context-aware*, C4=*Knowledge-guided*, and C5=*Compositions*) to the five GLUE tasks. For sentence-level perturbations, we apply *Syntactic-based perturbations* (C6) to the five GLUE tasks. *Distraction-based perturbations* (C7) are applied to four GLUE tasks without QQP, as they may affect the semantic similarity. For human-crafted examples, we apply *CheckList* (C8) to SST-2, QQP, and QNLI; *StressTest* (C9) and *ANLI* (C10) to MNLI; and *AdvSQuAD* (C11) to QNLI tasks.

Corpus	Task	Train (GLUE)	Test (AdvGLUE)	Word-Level					Sent.-Level		Human-Crafted			
				C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
SST-2	sentiment	67,349	1,420	204	197	91	175	64	211	320	158	0	0	0
QQP	paraphrase	363,846	422	42	151	17	35	75	37	0	65	0	0	0
QNLI	NLI/QA	104,743	968	73	139	71	98	72	159	219	80	0	0	57
RTE	NLI	2,490	304	43	44	31	27	23	48	88	0	0	0	0
MNLI	NLI	392,702	1,864	69	402	114	161	128	217	386	0	194	193	0
Sum of AdvGLUE test set			4,978	431	933	324	496	362	672	1013	303	194	193	57

<https://adversarialglue.github.io/>

<https://arxiv.org/pdf/2111.02840.pdf>

Evaluation on real-datasets

Table 1: Statistics of test sets in this paper

Area	Dataset	Task	#Sample	#Class
Adversarial robustness	SST-2	sentiment classification	148	2
	QQP	quora question pairs	78	3
	MNLI	multi-genre natural language inference	121	3
	QNLI	question-answering NLI	148	2
	RTE	textual entailment recognition	81	2
	ANLI	text classification	1200	3
	AdvGLUE-T	machine translation (En \rightarrow Zh)	30	-
OOD robustness	Flipkart	sentiment classification	331	2
	DDXPlus	medical diagnosis classification	100	50

Details

- Adv and OOD sentiment analysis:

RO is the sentiment of the following sentence positive or negative: The primitive force of this film seems to bybble up from the vast collective memory of the combatants.

The sentiment of the sentence is somewhat ambiguous and could be interpreted differently depending on the context and the reader's perspective. However, based on the use of the words "primitive force" and "combatants," it may suggest a somewhat negative or intense sentiment. The phrase "primitive force" implies something raw and unrefined, and the use of "combatants" suggests violence or conflict. However, the phrase "vast collective memory" could also be interpreted as positive or nostalgic. Overall, the sentiment of the sentence may be somewhat mixed, with elements of both positive and negative connotations.



RO Answer me with "positive" if the following sentence represents positive feelings or "negative" else: The primitive force of this film seems to bybble up from the vast collective memory of the combatants.

Negative.

Prompt=Answer me with "positive" if the following sentence represents positive feelings or "negative" else: +
XXX

OOD medical dataset

- "Imagine you are an intern doctor. Based on the previous dialogue, what is the diagnosis? Select one answer among the following lists: ['spontaneous pneumothorax', 'cluster headache', 'boerhaave', 'spontaneous rib fracture', 'gerd', 'hiv (initial infection)', 'anemia', 'viral pharyngitis', 'inguinal hernia', 'myasthenia gravis', 'whooping cough', 'anaphylaxis', 'epiglottitis', 'guillain-barré syndrome', 'acute laryngitis', 'croup', 'psvt', 'atrial fibrillation', 'bronchiectasis', 'allergic sinusitis', 'chagas', 'scombroid food poisoning', 'myocarditis', 'larygospasm', 'acute dystonic reactions', 'localized edema', 'sle', 'tuberculosis', 'unstable angina', 'stable angina', 'ebola', 'acute otitis media', 'panic attack', 'bronchospasm / acute asthma exacerbation', 'bronchitis', 'acute copd exacerbation / infection', 'pulmonary embolism', 'urti', 'influenza', 'pneumonia', 'acute rhinosinusitis', 'chronic rhinosinusitis', 'bronchiolitis', 'pulmonary neoplasm', 'possible nstemi / stemi', 'sarcoidosis', 'pancreatic neoplasm', 'acute pulmonary edema', 'pericarditis', 'cannot decide']. The answer should be a single word. The answer is: "

Details

- QQP




Is the following two questions equivalent? Just answer me with "yes" or "no":
Question 1: I am getting fat on my lower body and on the torso, is there any way I can get fit without looking skinny fat?
Question 2: Why I am getting skinny instead of losing body fat?



No.


Prompt: Is the following two questions equivalent? Just answer me with "yes" or "no": X




MNLI

 Tell me the relation between the following hypothesis and premise in one of the three category: entailment, contradiction, or neutral? Premise: You and your friends are not welcome here, said Severn.
Hypothesis: Severn said the people were not welcome there.

 The hypothesis and premise are in the category of entailment, where the hypothesis logically follows from the premise.  

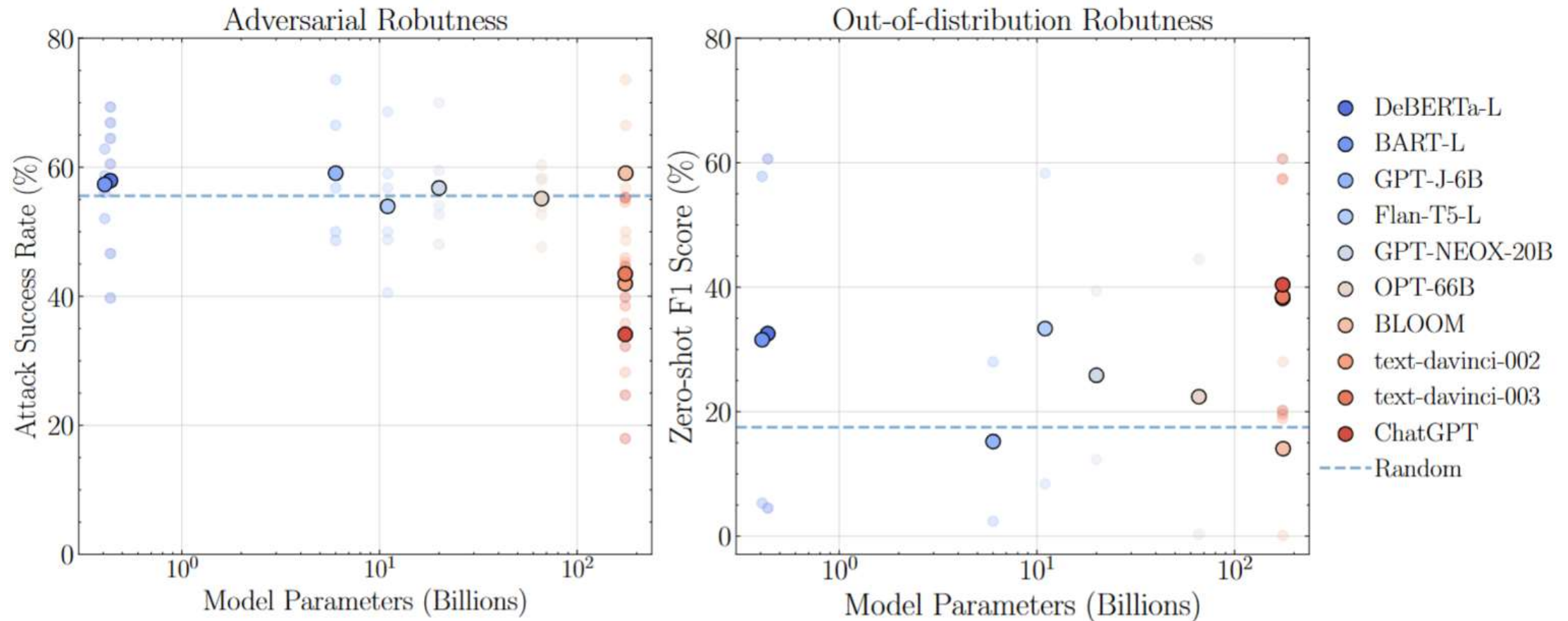
The hypothesis repeats the statement made by Severn in the premise with some minor changes in phrasing. Both the premise and hypothesis express the same idea that Severn said that "you and your friends" or "the people"

 Tell me the relation between the following hypothesis and premise in one of the three category: entailment, contradiction, or neutral? Answer me with one word.
Premise: You and your friends are not welcome here, said Severn.
Hypothesis: Severn said the people were not welcome there.

 Entailment.  

Prompt: Tell me the relation between the following hypothesis and premise in one of the three category: entailment, contradiction, or neutral? Answer me with one word.

Results: parameter vs. performance



Results

- ChatGPT achieves great performance
 - But still much room for improvement...
 - Overfitting? (DeBERTa-L vs. ChatGPT)

Table 2: Zero-shot classification results on adversarial (ASR↓) and OOD (F1↑) datasets. The best and second-best results are highlighted in **bold** and underline.

Model & #Param.	Adversarial robustness (ASR↓)						OOD robustness (F1↑)	
	SST-2	QQP	MNLI	QNLI	RTE	ANLI	Flipkart	DDXPlus
Random	50.0	50.0	66.7	50.0	50.0	66.7	20.0	4.0
DeBERTa-L (435 M)	66.9	39.7	64.5	46.6	60.5	69.3	60.6	4.5
BART-L (407 M)	56.1	62.8	58.7	52.0	56.8	<u>57.7</u>	57.8	5.3
GPT-J-6B (6 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	2.4
Flan-T5-L (11 B)	40.5	59.0	48.8	50.0	56.8	68.6	58.3	8.4
GPT-NEOX-20B (20 B)	52.7	56.4	59.5	54.0	48.1	70.0	39.4	12.3
OPT-66B (66 B)	47.6	53.9	60.3	52.7	58.0	58.3	44.5	0.3
BLOOM (176 B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	0.1
text-davinci-002 (175 B)	46.0	<u>28.2</u>	54.6	45.3	35.8	68.8	57.5	18.9
text-davinci-003 (175 B)	44.6	55.1	<u>44.6</u>	<u>38.5</u>	<u>34.6</u>	62.9	57.3	<u>19.6</u>
ChatGPT (175 B)	<u>39.9</u>	18.0	32.2	34.5	24.7	55.3	60.6	20.2

Zero-shot classification

Model	BLEU↑	GLEU↑	METOR↑
OPUS-MT-EN-ZH	18.11	26.78	46.38
Trans-OPUS-MT-EN-ZH	15.23	24.89	45.02
text-davinci-002	24.97	36.30	<u>59.28</u>
text-davinci-003	30.60	40.01	61.88
ChatGPT	<u>26.27</u>	<u>37.29</u>	58.95

Machine translation

Case study: adv. examples

Table 4: Case study on adversarial examples. Adversarial manipulations are marked **red**.

Type	Input	Truth	davinci003	ChatGPT
word-level (typo)	i think you 're here for raunchy college humor .	Positive	Negative	Negative
	Mr. Tsai is a very original artist in his medium , and what time is it there?	Positive	Positive	Positive
	Q1: Can you TRANSLATE these to English language? Q2: Cn you translate ths from Bengali to English lagnuage ?	Not equivalent	Not equivalent	Equivalent
	Q1: What are the best things in Hog Kong? Q2: What is the best thing in Hong Kong?	Equivalent	Not equivalent	Not equivalent
sentence-level (distraction)	Question: What is the minimum required if you want to teach in Canada? Sentence: @KMcYo0 In most provinces a second Bachelor's Degree such as a Bachelor of Education is required to become a qualified teacher.	Not entailment	Entailment	Entailment
	Question: @uN66rN What kind of water body is rumored to be obscuring Genghis Khan's burial site? Sentence: Folklore says that a river was diverted over his grave to make it impossible to find (the same manner of burial as the Sumerian King Gilgamesh of Uruk and Atilla the Hun).	Entailment	Not entailment	Not entailment
	https://t.co/1GPp0U the iditarod lasts for days - this just felt like it did .	Negative	Positive	Negative
	holden caulfield did it better . https://t.co/g4vJKP	Negative	Positive	Negative

Case study: OOD examples

- It's hard to find OOD examples for large models...

Table 6: Case study on OOD examples.

Input	Truth	davinci003	ChatGPT
quality of cover is not upto mark but the content in the book is really good from foundation to difficult level questions are of latest pattern great work	Positive	Positive	Positive
worst product dont buy flipcart should not sell such useless product prepared food only one time it damaged smoke came out and burned it good for nothing	Positive	Negative	Negative
definitely it will not fit wagon r either front or back it will cover one side fully and the other side partially thickness is not that much average product	Positive	Negative	Negative
this ink is genuine but the problem with printer is it shows red light after 100pages but i still used the cartridge and at last 357 pages were printed	Negative	Positive	Neutral
working fine good but received in messy box and there is bent on inverter at corner think mistake of courier facility whatever but working fine no issue	Negative	Positive	Positive

Back to vision models

- ViT-22B by Google
 - By far the largest vision foundation models?

Table 3: Zero-shot transfer results on ImageNet (variants).

Model	IN	IN-v2	IN-R	IN-A	ObjNet	ReaL
CLIP	76.2	70.1	88.9	77.2	72.3	-
ALIGN	76.4	70.1	92.2	75.8	72.2	-
BASIC	85.7	80.6	95.7	85.6	78.9	-
CoCa	86.3	80.7	96.5	90.2	82.7	-
LiT-g/14	85.2	79.8	94.9	81.8	82.5	88.6
LiT-e/14	85.4	80.6	96.1	88.0	84.9	88.4
LiT-22B	85.9	80.9	96.0	90.1	87.6	88.6

Vision is not solved, for now...

<https://arxiv.org/pdf/2302.05442.pdf>

<https://github.com/jindongwang/transferlearning/tree/master/code/clip>

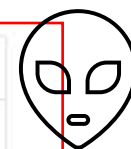
DomainNet

backbone	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	avg
RN50	0.5158	0.3920	0.5281	0.0627	0.7688	0.4886	0.4593
RN101	0.5981	0.4070	0.5676	0.1030	0.7935	0.5417	0.5018
RN50x4	0.6335	0.461	0.6131	0.1001	0.8115	0.5799	0.5332
RN50x16	0.6876	0.4715	0.6351	0.1266	0.8232	0.6301	0.5624
RN50x64	0.7328	0.5024	0.6763	0.1626	0.8463	0.6749	0.5992
ViT-B-32	0.6703	0.3992	0.6239	0.1318	0.8054	0.5853	0.5360
ViT-B-16	0.7091	0.4679	0.6599	0.1442	0.8315	0.6343	0.5745
ViT-L-14	0.7795	0.4958	0.6913	0.2247	0.8599	0.7023	0.6256
ViT-L-14@336px	0.7860	0.5226	0.7078	0.2231	0.8662	0.7163	0.6370

CLIP

TerraInc

backbone	L38	L43	L46	L100	avg
RN50	0.1361	0.3297	0.2169	0.0884	0.1928
RN101	0.4197	0.3748	0.2674	0.2474	0.3273
RN50x4	0.2626	0.3567	0.2438	0.3164	0.2949
RN50x16	0.3532	0.4715	0.3427	0.3626	0.3825
RN50x64	0.4083	0.4990	0.3672	0.5817	0.4641
ViT-B-32	0.1339	0.3071	0.1844	0.1346	0.1900
ViT-B-16	0.1958	0.3350	0.3165	0.5117	0.3398
ViT-L-14	0.4008	0.4597	0.3760	0.5182	0.4387
ViT-L-14@336px	0.4295	0.4892	0.4071	0.5100	0.4590



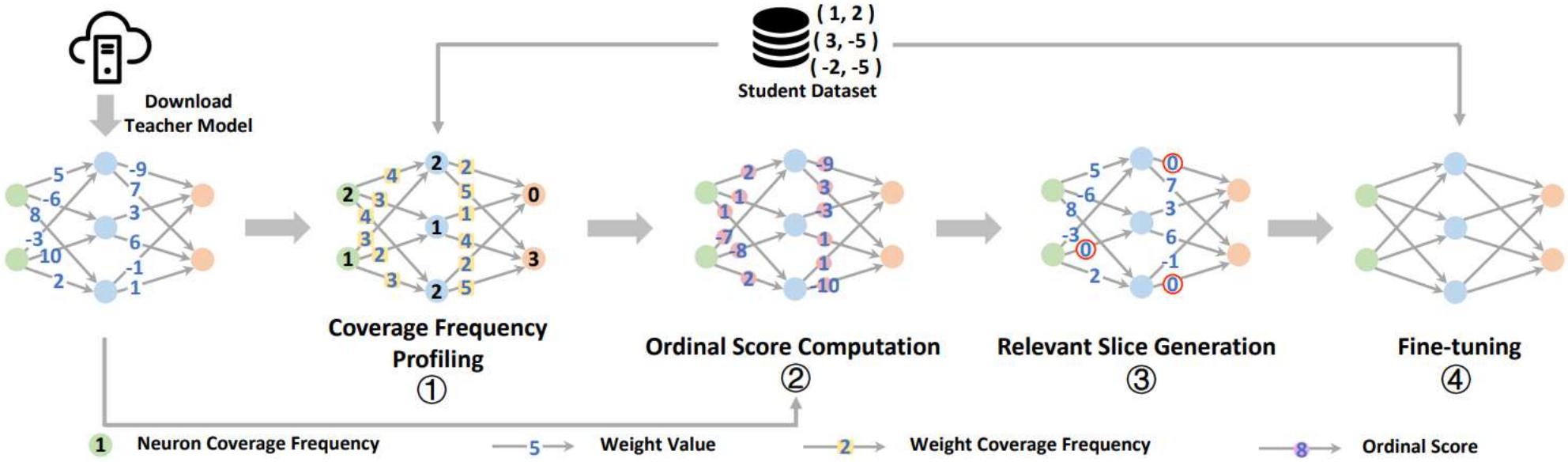
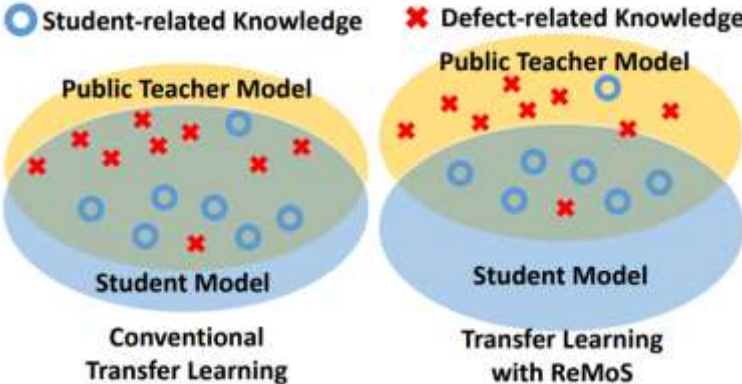
Possible directions

- Data cleaning
 - A balanced: what is noise and what is not?
- Spell correction
 - Training a spell-free NLP model?
- Adversarial defense
 - A robust pre-trained model? Or a robust fine-tuning technology?
 - Selective forgetting of malicious weights: [Zhang et al, 2022]
- Large models + OOD algorithms?
 - Efficient! [Lu et al. 2023]

- Zhang et al. ReMoS: Reducing Defect Inheritance in Transfer Learning via Relevant Model Slicing. ICSE 2022.
- Lu et al. FedCLIP: Fast Generalization and Personalization for CLIP in Federated Learning.

ReMoS: relevant model slicing

- Select the most important weights and then forget all others
 - Safe fine-tuning



• Zhang et al. ReMoS: Reducing Defect Inheritance in Transfer Learning via Relevant Model Slicing. ICSE 2022.

Conclusions

- Adv. And OOD robustness remain a major challenge
 - The best performance brought by ChatGPT
 - But still far from perfection
- OOD is NOT solved by LLMs
 - Domain-specific finetuning is still needed



Thanks for listening!

Contact: Jindong.wang@microsoft.com
<https://jd92.wang/>